Received: 12 December 2019 | Revised: 8 June 2020 | Accepted: 25 August 2020

DOI: 10.1002/pds.5121

WILEY

ORIGINAL REPORT

Implementing high-dimensional propensity score principles to improve confounder adjustment in UK electronic health records

John Tazare[1] | Liam Smeeth[1,2] | Stephen J. W. Evans[1] | Elizabeth Williamson[1,2] | Ian J. Douglas[1,2]

[1]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

[2]Health Data Research UK, London, UK

Correspondence
John Tazare, Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, Keppel St, Bloomsbury, London WC1E 7HT, UK.
Email: john.tazare1@lshtm.ac.uk

Funding information
Medical Research Council, Grant/Award Numbers: MR/N013638/1, MR/M013278/1, MR/S01442X/1

Abstract

Purpose: Recent evidence from US claims data suggests use of high-dimensional propensity score (hd-PS) methods improve adjustment for confounding in non-randomised studies of interventions. However, it is unclear how best to apply hd-PS principles outside their original setting, given important differences between claims data and electronic health records (EHRs). We aimed to implement the hd-PS in the setting of United Kingdom (UK) EHRs.

Methods: We studied the interaction between clopidogrel and proton pump inhibitors (PPIs). Whilst previous observational studies suggested an interaction (with reduced effect of clopidogrel), case-only, genetic and randomised trial approaches showed no interaction, strongly suggesting the original observational findings were subject to confounding. We derived a cohort of clopidogrel users from the UK Clinical Practice Research Datalink linked with the Myocardial Ischaemia National Audit Project. Analyses estimated the hazard ratio (HR) for myocardial infarction (MI) comparing PPI users with non-users using a Cox model adjusting for confounders. To reflect unique characteristics of UK EHRs, we varied the application of hd-PS principles including the level of grouping within coding systems and adapting the assessment of code recurrence. Results were compared with traditional analyses.

Results: Twenty-four thousand four hundred and seventy-one patients took clopidogrel, of whom 9111 were prescribed a PPI. Traditional PS approaches obtained a HR for the association between PPI use and MI of 1.17 (95% CI: 1.00-1.35). Applying hd-PS modifications resulted in estimates closer to the expected null (HR 1.00; 95% CI: 0.78-1.28).

Conclusions: hd-PS provided improved adjustment for confounding compared with other approaches, suggesting hd-PS can be usefully applied in UK EHRs.

KEYWORDS

confounder adjustment, database research, electronic health records, electronic medical records, high-dimensional propensity score, pharmacoepidemiology

This work was presented as a poster at 35th ICPE Philadelphia 2019.

1 | INTRODUCTION

Electronic Health Records (EHRs) are increasingly used for research investigating the effects of medications.[1,2] Adequate adjustment for confounding remains a key issue and incorrect conclusions can be drawn amid concerns of residual or unmeasured confounding.[3,4]

Developed in US claims data to improve confounder adjustment, the high-dimensional propensity score (hd-PS) approach treats information stored within healthcare databases as proxies for key underlying confounders.[5] Some proxies may be strongly correlated with variables typically included in a traditional propensity score (PS) analysis; others may represent information about patients that is otherwise unmeasured, for example, frailty.[5]

Despite application in various settings (including UK EHRs),[6-9] detailed guidance on how to apply the hd-PS outside US claims data is lacking. Important differences between data sources mean that careful consideration is needed when implementing hd-PS principles to ensure source-specific characteristics are handled appropriately.

We propose a series of modifications to the hd-PS that aim to characterise key features of UK EHRs whilst adhering to the underlying principles.[5,6]

KEY POINTS

1. High-dimensional propensity score (hd-PS) approaches are a popular method for confounder adjustment in healthcare databases.
2. Whilst the performance of hd-PS is well established in US claims data, there is a lack of guidance for applying hd-PS principles in other settings.
3. We propose modifications to better tailor the hd-PS to UK electronic health records and apply these to a recent cohort study where results strongly suggested residual confounding.
4. The modified hd-PS achieved results closer to those obtained by a randomised controlled trial.
5. We have demonstrated that hd-PS approaches can be usefully applied in UK electronic health records to achieve improved confounder adjustment.

2 | PROPENSITY SCORES

The PS is the conditional probability of being treated given a set of observed covariates.[10-12]

PSs model the treatment allocation process and therefore offer advantages over multivariable analysis in EHRs, since investigators are forced to consider indications for treatment use and can convert large amounts of confounder information into a single number.[4]

At a particular value of the PS, the distribution of observed covariates is balanced between treated and untreated individuals, allowing consistent estimation of treatment effects, assuming all confounders are included in the model.[13]

3 | DESCRIPTION OF THE hd-PS APPROACH AND UNDERLYING PRINCIPLES

3.1 | Preliminary steps

Demographics (*d*) and clinical factors believed to be important confounders (*l*) are forced into the PS model.[5] A baseline time-window for assessing patient confounder information is established (often 1 year before study entry date).

3.2 | Identification of most relevant covariates

Relevant information in the database is separated into *p* dimensions.[5] The underlying principle is that each dimension should represent a different aspect of care relevant to the healthcare system under investigation (principle 1). For example, in US claims data, it is typical to separate information pertaining to diagnoses, procedures and prescribing.[5]

Healthcare databases typically store information in the form of thousands of discrete codes which vary by database. To avoid sparsity, information is often grouped at a granularity level set by the investigator that captures related aspects of health status and care (principle 2). We illustrate this using an example from the International Classification of Diseases (ICD-10).[14] The ICD-10 coding system is hierarchical meaning that all information pertaining to one concept, for example type 2 diabetes mellitus (T2DM), begins with the same 3-character code (E11 for T2DM).

Code groups are ranked by prevalence and investigators prespecify a number to be selected from each dimension.[5]

Code frequency is then assessed for each individual; measuring the recurrence of identified codes in the baseline time-window. This is summarised by three indicator variables:

Once: Code is recorded ≥ once.

Sporadic: Code is recorded ≥ the median

Frequent: Code is recorded ≥ the 75th percentile

This classification assumes that frequency of recording relates to the importance of a code as a descriptor a patient's health status (principle 3).

3.3 | Prioritisation

The steps so far generate a large pool of potential confounders. Attempting to include all of these variables in the PS model would often lead to concerns of overfitting therefore a variable selection step is necessary to ensure statistical stability.

The hd-PS uses the Bross formula to prioritise covariates across dimensions by their potential to bias the treatment-outcome relationship.[5,15,16] This has three components. Firstly, it takes the confounded apparent relative risk (ARR) for a particular binary covariate as a function of the relative risk (RR) in the absence of confounding by this covariate. Secondly, the imbalance in prevalence amongst the exposed (P_{C1}) and unexposed (P_{C0}) patients. Thirdly, the independent association between a confounder and the study outcome (RR_{CD}):

$$ARR = RR \times bias_M$$

$$\text{where } bias_M = \frac{P_{C1}(RR_{CD}-1)+1}{P_{C0}(RR_{CD}-1)+1} \text{for all } RR_{CD}.$$

Each dimension is sorted in descending order by the magnitude of $|\log(bias_M)|$. This bias term takes a larger value the greater the potential a covariate has to bias the relationship of interest. Therefore, the top *k* empirical covariates are included in the PS. Typically several hundred covariates are selected.

3.4 | Estimation of the hd-PS

The selected empirical covariates are added to the predefined variables before estimating the PS. Traditional PS methods are then used to estimate the treatment effect.[12] The final principle is that after accounting for the top *k* empirically selected covariates, residual confounding effects are assumed to be negligible (principle 4).

4 | PROPOSED IMPLEMENTATION OF hd-PS PRINCIPLES TO UK EHRS

In this section, issues surrounding the translation of hd-PS principles to UK EHRs are discussed alongside our proposed modifications (summarised in Figure 1).

4.1 | Principle 1: Identification of dimensions

There are important differences between insurance claims and EHR data in terms of data availability, structure and the reasons for data recording.[17,18] This necessitated the identification of clinically relevant dimensions based on patient contact with primary care services in the UK. Since previous applications of hd-PS in UK EHRs have not reached a consensus about what these dimensions should be, we drew on general practitioner (GP) experience within our research team.[9,19] We proposed three dimensions separating clinical, referral and prescription information (summarised in Table 1).

4.2 | Principle 2: Code granularity

Data in the clinical and referral dimension are recorded using the Read code system.[20] Read codes are less structured than coding systems used in claims databases (eg, ICD-10[14]). Consequently, the Read coding system does not fully capture distinct concepts at any level of granularity. For example, whilst the Read code 1434.00 relates to

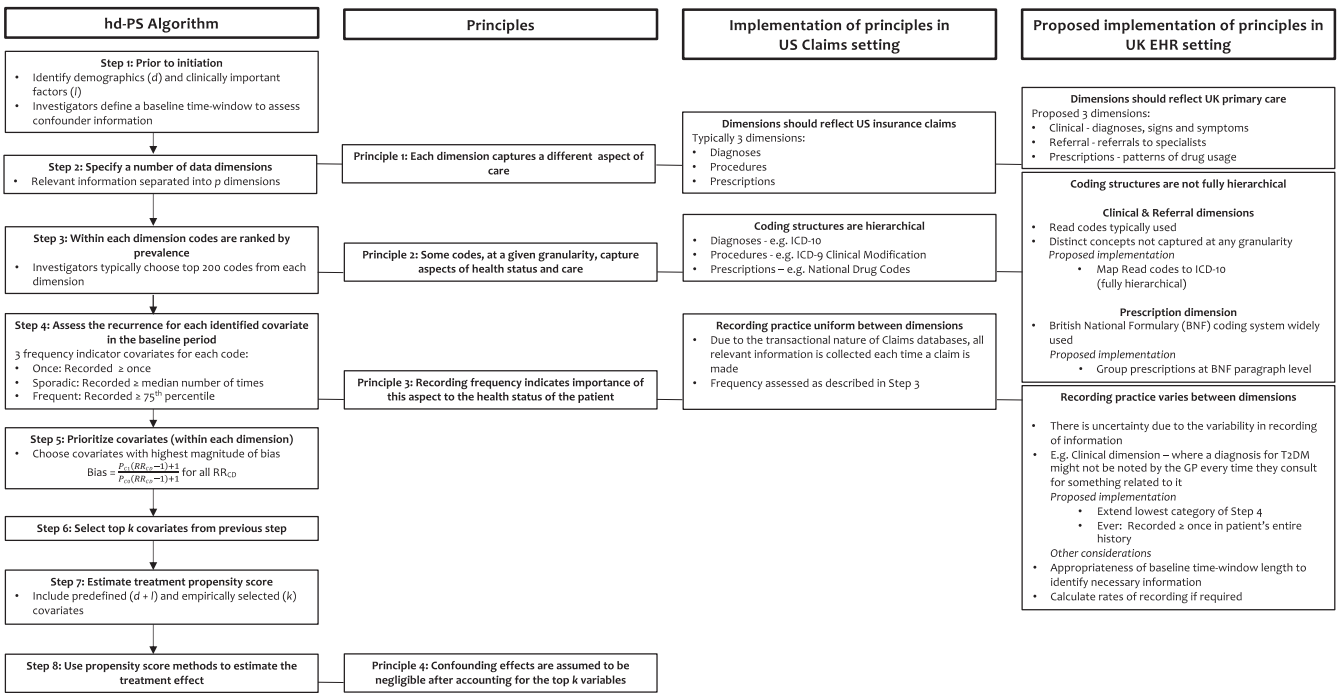

FIGURE 1 Flowchart depicting hd-PS steps, underlying principles and adaptations for translating to UK electronic health records. GP, general practitioner; hd-PS, high-dimensional propensity score; ICD-10, International Classification of Disease; T2DM, type 2 diabetes mellitus

TABLE 1 Summary of dimensions for UK electronic health recordss

Dimension	Information included	Health status and care
Clinical	Diagnoses, signs and symptoms[a]	Indicates underlying health of patient and frequency of contact with healthcare system
Referral	Referrals to specialists	Indicates escalation in care or investigation
Prescriptions	Drug prescriptions issued in primary care	Frequency and patterns of drug usage

[a]The clinical dimension also contains information relating to administrative codes or references to measurements that occurred without results.

history of diabetes mellitus, grouping codes at the three-digit level (eg, 143) would capture concepts in addition to diabetes such as codes relating to thyroid disorder. Therefore, two codes with the same three-digit Read code may capture disparate clinical concepts, whereas conversely, two codes capturing similar concepts may have different three-digit Read codes.

A manual solution to group all Read codes at a level capturing distinct medical concepts is not practical, therefore we mapped Read codes to the ICD-10 coding system. This was achieved using cross maps developed by NHS Digital[21] and allowed replication of the approach taken by Schneeweiss et al,[5] which hierarchically grouped distinct medical concepts at a certain granularity level.

For the prescription dimension the British National Formulary (BNF) coding system is used. We classified prescriptions at the BNF paragraph level which typically groups prescriptions by indication rather than mechanism of action.[22]

4.3 | Principle 3: Code recurrence

Code frequency is assessed by the hd-PS to provide an indicator of a patient's underlying health.[5] In claims data all relevant information is recorded at each instance a claim is completed which leads to an intrinsic link between disease severity and code frequency.

EHRs exist for clinical record keeping which means that such a link is harder to discern since all relevant information will not necessarily be recorded at each consultation. Frequency of recording is instead likely to be a function of several factors including severity of illness, frequency of consultation and GP preference.

We classified the frequency of codes in a pre-specified baseline time-window, 1 year prior to study entry. Recognising the variability in recording we replaced the "Once" indicator with an "Ever" indicator which captured whether a code had been recorded during a patient's entire history. The remaining frequency indicators were assessed during the baseline time-window.

We hypothesised that the degree to which information is recorded at each consultation was likely to vary by dimension, with more complete recording likely in the prescription and referral dimensions. However, in the clinical dimension relevant information is often not re-recorded at each consultation. For example, a patient receiving prescriptions relating to a diagnosis of T2DM will have this diagnosis recorded but not necessarily at each relevant consultation.

To investigate whether this information was likely to be overlooked when assessing information in a narrow time-window we extended the baseline time-window for the Clinical dimension. Acknowledging the fact that patients will have varying lengths of baseline information available we classified the frequency of codes by assessing rates in instead of counts. We used three indicators to classify our revised frequency assessment (see Figure 1 for full definition).

4.4 | Principle 4: Selected number of variables

The capacity of the hd-PS to control for confounding can be sensitive to the number of covariates selected.[23,24] Whilst in claims data investigators typically specify 500 empirical covariates it is unclear if this is appropriate in UK EHRs. We investigated the impact of selecting 100, 250, 500 and 750 covariates.

5 | APPLICATION TO EXAMPLE IN CPRD

5.1 | Data

The Clinical Practice Research Datalink (CPRD) is a de-identified primary care database broadly representative of patients registered at GPs in the UK. It includes data pertaining to prescribing, diagnosis, referrals and some lifestyle factors for approximately 9% of the UK population.[20]

A recent cohort study using the CPRD linked with the Myocardial Ischaemia National Audit Project (MINAP) investigated the combined use of proton pump inhibitors (PPI) with clopidogrel and aspirin. A possible interaction whereby PPIs may reduce the conversion of clopidogrel to its active metabolite had been suggested, raising concerns that combined use may lead to a reduction in clopidogrel effectiveness and an increased risk of vascular events. The cohort analysis found that combined use was indeed associated with an increased risk of myocardial infarction (MI).[3]

The pattern of associations found strongly suggested that residual confounding between patients may have explained the results as they were not specific to MI and were found for both strong and weak inhibitors of cytochrome P450 3A4 (the mechanism proposed for the drug interaction). Furthermore, a self-controlled case series (SCCS) analysis[25] conducted on the same data found no evidence of increased risk.

The authors concluded that the results from the cohort study reflect confounding in the cohort estimate. In addition, unconfounded studies based on genetic instrumental variable approaches using genetic effects on drug metabolism pathways also suggested no evidence of increased risk.[26] A randomised double-blind trial has

subsequently also suggested a lack of clinical effect of PPIs on MI risk, when used in combination with clopidogrel (HR = 0.92; 95% CI: 0.44-1.90).[27]

5.2 | Design

We summarise the original study design conducted by Douglas et al.[3] Patients had to be present in the CPRD with at least 12 months of prior registration before first prescription for clopidogrel. Study entry was defined as the latest of first recorded clopidogrel prescription in combination with aspirin or 1 January 2003. Patients were censored at the earliest of stopping treatment for aspirin or clopidogrel, death, transferring out of the practice, last data collection date for the practice, 31 July 2009 or an occurrence of MI. Exposure was defined as any prescription for a PPI. We focus on the incident MI outcome which was ascertained using the MINAP database.

5.3 | Statistical analysis

The original study analysed the hazard ratio (HR) for the association between PPI treatment and MI using Cox models, adjusting for 14 selected confounders. Missing data for body mass index, smoking and alcohol consumption were handled using missing categories. These conditions were applied consistently across all analyses.

We reanalysed the original data taking an intent-to-treat approach that classified patients according to original exposure status and incorporated baseline confounder information using PSs. We estimated the PS using multivariable logistic regression to model the relationship between treatment and potential confounders. Inverse probability of treatment weights (IPTW) were calculated from the PS which essentially constructs two synthetic samples representing the scenarios in which everyone had been treated and everyone had been untreated.[11] A weighted Cox model incorporating the IPTWs was used to model the outcome.

Unless otherwise stated, all hd-PS analyses defined the three aforementioned dimensions and assessed patient confounder information recorded in the year prior to cohort entry. The top 200 most prevalent codes were selected from each dimension and 500 covariates were included in the PS model.

We performed a standard hd-PS analysis which implemented the algorithm using Read codes (classified at three-character Read code granularity) for the clinical and referral dimensions. All Read codes were included regardless of whether they map to ICD-10 to represent the default position of applying the method wholesale to the coded data in these dimensions. We then applied our modifications: mapping the clinical and referral dimensions to ICD-10 and extending the frequency assessment.

A sensitivity analysis extended the baseline time-window to 3 and 5 years for the Clinical dimension. We also investigated the impact of selecting 100, 250 and 750 covariates on confounding control.

All HR results are presented with 95% confidence intervals in parentheses. Analyses were conducted using Stata 14.[28]

6 | RESULTS

Demographics and clinical characteristics for the cohort study are summarised in Table 2. Twenty-four thousand four hundred and seventy-one patients took clopidogrel, of whom 9111 were prescribed a PPI. Of PPI users, 313 (3.4%) had an incident MI vs 421 (2.7%) in the non-users. Users of PPIs were older and were more likely to have had a history of cancer, diabetes or peripheral vascular disease compared to non-users (Table 2).

TABLE 2 Baseline characteristics by proton pump inhibitor status amongst clopidogrel and aspirin users

	Clopidogrel and aspirin users			
	No PPI		PPI	
	N = 15 360		N = 9111	
Demographics	N	(%)	N	(%)
Median age (years)	68.9		71.1	
Sex				
Male	10 007	(65.1)	5323	(58.4)
Body mass index (kg/m^2)				
<20	480	(3.1)	429	(4.7)
20-25	3987	(26.0)	2339	(25.7)
>25	10 004	(65.1)	5809	(63.8)
Missing	889	(5.8)	534	(5.9)
Smoking status				
Non-smoker	4781	(31.1)	2780	(30.5)
Current	2760	(18.0)	1503	(16.5)
Ex-smoker	7777	(50.6)	4799	(52.7)
Missing	42	(0.3)	29	(0.3)
Alcohol status				
Non-drinker	1528	(9.9)	1080	(11.9)
Ex-drinker	938	(6.1)	687	(7.5)
Amount not specified	399	(2.6)	254	(2.8)
<2 units/d	3060	(19.9)	1908	(20.9)
3–6 units/day	7488	(48.8)	4106	(45.1)
>6 units/d	1180	(7.7)	606	(6.7)
Status unknown	767	(5.0)	470	(5.2)
History of				
Diabetes	4404	(28.7)	3090	(33.9)
Peripheral vascular disease	1629	(10.6)	1095	(12.0)
Coronary heart disease	12 198	(79.4)	7292	(80.0)
Ischaemic stroke	1571	(10.2)	954	(10.5)
Cancer	2038	(13.3)	1381	(15.2)

Abbreviation: PPI, proton pump inhibitor.

TABLE 3 Estimated treatment effect of proton pump inhibitor use on myocardial infarction risk by variations in high-dimensional propensity score approach

Model	Dimension code granularity	Baseline assessment period	Most prevalent codes selected by dimension	Code frequency assessment	Covariates included in propensity score model	Total covariates in propensity score model	Outcome model HR (95% CI)	log(HR) SE
1	…	…	…	…	Unadjusted	…	1.23 (1.06 to 1.42)	0.08
2	…	…	…	…	Demographics + predefined[a]	*d* = 2, *l* = 8	1.17 (1.00 to 1.35)	0.10
3	3-digit Read[b] + BNF[c]	1 year	200	Counts	+ Empirical covariates	*d* = 2, *l* = 8, *k* = 500	1.07 (0.86to 1.34)	0.11
4	3-digit ICD-10[d] + BNF	1 year	200	Counts	+ Empirical covariates	*d* = 2, *l* = 8, *k* = 500	1.15 (0.91 to 1.45)	0.12
5	3-digit ICD-10 + BNF	1 year	200	Ever category + counts	+ Empirical covariates	*d* = 2, *l* = 8, *k* = 500	1.00 (0.78 to 1.28)	0.13
6	3-digit ICD-10 + BNF	3 years	200	Ever category + counts + rates (clinical dimension)	+ Empirical covariates	*d* = 2, *l* = 8, *k* = 500	1.12 (0.91 to 1.39)	0.11
7	3-digit ICD-10 + BNF	5 years	200	Ever category + counts + rates (clinical dimension)	+ Empirical covariates	d = 2, l = 8, k = 500	1.10 (0.90 to 1.36)	0.11
8	3-digit ICD-10 + BNF	1 year	200	Ever category + counts)	+ Empirical covariates	*d* = 2, *l* = 8, *k* = 100	1.07 (0.87 to 1.32)	0.10
9	3-digit ICD-10 + BNF	1 year	200	Ever category + counts	+ Empirical covariates	*d* = 2, *l* = 8, *k* = 250	1.02 (0.81 to 1.27)	0.12
10	3-digit ICD-10 + BNF	1 year	200	Ever category + counts	+ Empirical covariates	*d* = 2, *l* = 8, *k* = 750	1.03 (0.79 to 1.28)	0.13

Abbreviations: *d*, number of demographics; *k*, number of variables empirically selected by the algorithm; *l*, number of predefined covariates.

[a]Demographics: age, sex; predefined covariates: smoking status, alcohol status, categorised BMI, alcohol status, history of PVD, CHD, stroke, cancer.

[b]Clinical terms are defined using Read codes in the Clinical Practice Research Datalink.

[c]British National Formulary (BNF) code at paragraph level.

[d]International Classification of Disease (ICD-10).

For the modified analyses, we mapped the clinical and referral dimensions from Read code to ICD-10. A large number of Read codes represent non-clinical information, for example, codes relating to administrative procedures. Since the aim of the mapping procedure is solely to capture clinically relevant information unmapped Read codes were expected. Upon inspection, the resulting unmapped codes could generally be categorised as either administrative information (eg, a letter), an indicator of a completed test without the result (eg, "blood pressure reading was taken") or coarse information we would typically include more granularly in the pre-defined covariates (eg, broad smoking terms). We include a sample of the most frequently occurring unmapped Read codes in the Supporting Information.

Results for all analyses are presented in Table 3. Using the confounders originally identified by Douglas et al[3] we obtained a HR for the association between PPI use and MI of 1.17 (1.00-1.35).

Applying our modifications reduced the HR for the association between PPI use and MI moving it towards a null result (Figure 2). The fully modified hd-PS obtained an HR of 1.00 (0.78 to 1.28).

In sensitivity analyses, extending the baseline time-window for the Clinical dimension lead to point estimates further from the null. Varying the number of covariates did not meaningfully alter point estimates. However, selecting fewer than 500 variables did improve the precision of effect estimates (Table 3).

We investigated the estimated PS distributions by treatment group obtained from investigator led and hd-PS analyses (Figure 3). These distributions compare the characteristics of patients in the populations under investigation. Compared to the investigator led approach, the hd-PS exposed greater variation between the treatment groups and captured extra predictors of prescribing which were also causing confounding bias.

7 | DISCUSSION

In this study, we aimed to optimise the application of hd-PS principles in UK EHR data. To investigate the potential of the hd-PS to account for residual confounding we took a study where the authors were confident the result obtained was subject to strong between patient confounding. We aimed to get an improved point estimate, closer to the expected null result, with similar precision to the original study. After mapping Read to ICD-10 codes, changing the frequency assessment, selecting 500 variables for inclusion and having a 1 year assessment period for covariates, our final hd-PS model obtained an HR for the association between MI and PPI use of 1.00 (0.78-1.28), compared to 1.17 (1.00-1.35) using confounders selected using an investigator led approach. Our modifications therefore achieved results closer to those obtained by a randomised double-blind trial, although the precision does not rule out results obtained from other studies.[3,27] Sensitivity analyses suggested that extending the covariate assessment period for the Clinical dimension to 3 or 5 years might not be helpful in this setting.

The authors of the original study had suspected unmeasured frailty or comorbidity severity was different between PPI users and non-users. Here, we have demonstrated that differences between PPI users and non-users are more apparent when using hd-PS than with traditional approaches. This highlights the potential for hd-PS approaches to include proxies for influential but unmeasured information regarding a patient's underlying health status.

Our adaptations aimed to tailor the hd-PS to UK EHRs and should be considered when applying the hd-PS in UK EHR data. The mapping of clinical and referral information to ICD-10 allows for the identification of homogeneous clinically meaningful proxies to be included in the hd-PS, although we acknowledge that information contained in

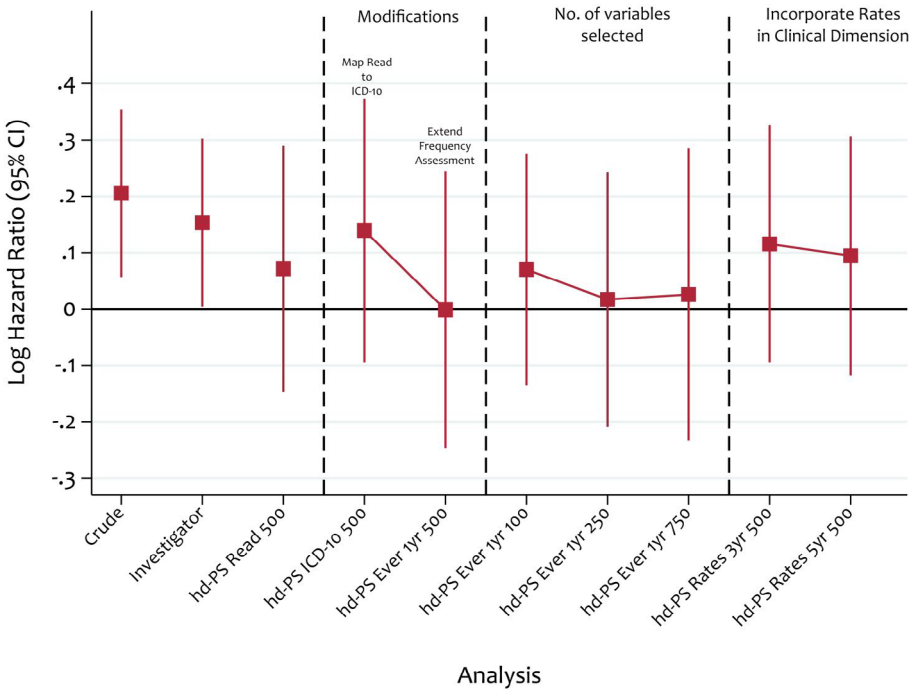

FIGURE 2 Empirical performance of hd-PS across our implemented adaptations. hd-PS, high-dimensional propensity score; ICD-10, International Classification of Disease

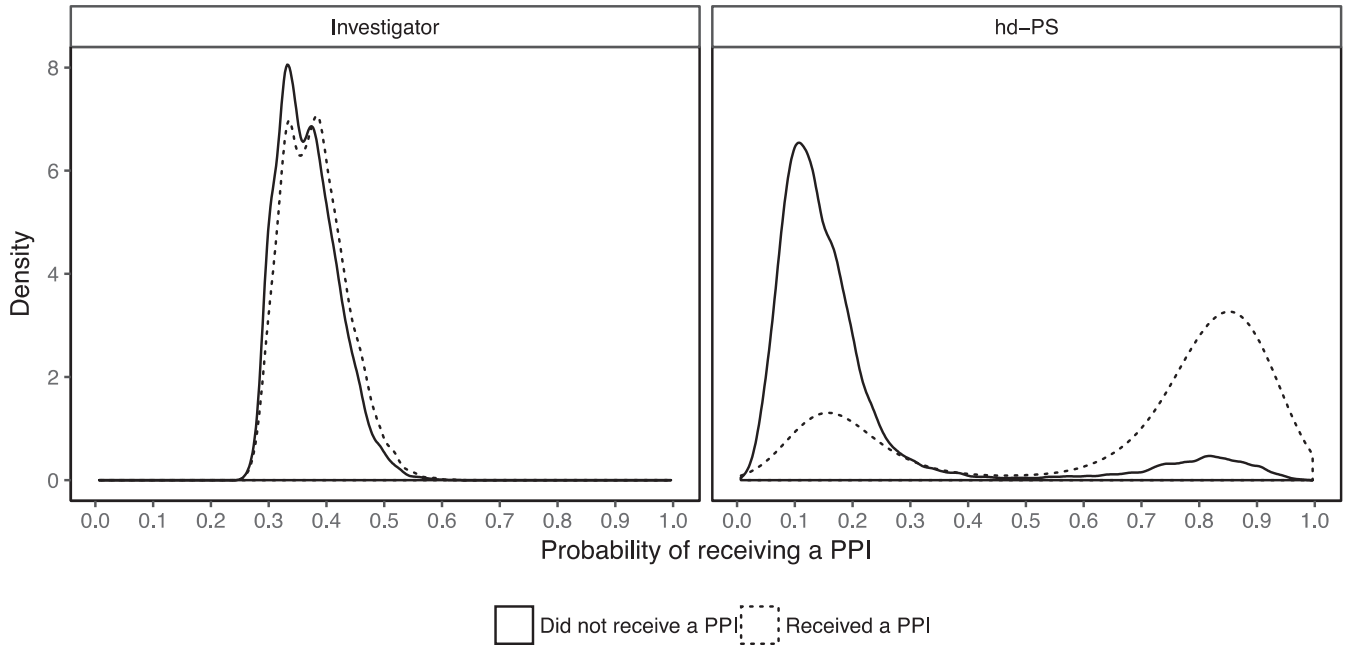

FIGURE 3 Comparison of the estimated propensity score from investigator and hd-PS approaches. hd-PS, high-dimensional propensity score; PPI, proton pump inhibitor

the unmapped codes is lost in this process. The inclusion of an Ever category to the frequency assessment of the hd-PS also more accurately captures recording practice in EHRs. Selecting 500 variables for inclusion in the final hd-PS model performed well, however selecting fewer variables obtained a very similar result with improved precision. The framework we have built could also be extended to include laboratory test results and free text information, the latter of which has been previously explored.[6]

Whilst there have been several developments to the hd-PS since its inception,[6] there has been little exploration of how to translate the algorithm beyond claims data. Much of this development work for hd-PS has been focussed on demonstrating it obtains known associations, such as the effect of non-steroidal anti-inflammatory drugs on the risk of gastrointestinal bleed.[5,9,24,29] However, these results have also been obtained through traditional methods of confounder adjustment. In the case study we present, a hd-PS approach has removed a known confounded association discovered using traditional methods.

Future applications of the hd-PS in this context will benefit from updates to the cross-map between Read and ICD-10. In the literature accompanying these cross-maps NHS Digital state that not every concept in one coding system can or should be represented in another.[21] NHS Digital's intention was to map clinically meaningful terms only, and it was reassuring to observe that the majority of unmapped Read codes were clinically uninformative and would typically be discarded in an investigator analysis (see Supporting Information).

When calculating the SEs for treatment effects we have ignored variable selection or estimation of the PS. Theoretically, this is likely to result in narrower confidence intervals,[30] although the practical consequences are yet to be fully explored. We obtained a bias-corrected bootstrap 95% CI based on 1000 replications for our final model of 0.70 to 1.30 (final model: [HR = 1.00; 95% CI: 0.78-1.28]).

Our results highlight the potential benefit of employing hd-PS approaches in EHR studies, especially to overcome intractable confounding. However, the hd-PS is not a panacea and we acknowledge that in studies where the confounding structure is relatively simple, the robustness of results is unlikely to differ between traditional and hd-PS methods. We recognise the need for further exploration of the hd-PS in this setting, via both controlled conditions and case studies. One outstanding issue surrounds the transparency of reporting when using hd-PS approaches and there is a need for tools to better communicate proxies included in the final hd-PS model.

This study has shown that the application of hd-PS methods outside the context of claims data requires careful consideration of how to optimally apply hd-PS principles. By adapting hd-PS principles to the UK EHR setting we have demonstrated the potential for hd-PS to improve confounder adjustment in EHRs.

CONFLICT OF INTEREST

I.J.D. reports grants from GlaxoSmithKline, ABPI and NIHR for projects unrelated to the submitted work and shares in GlaxoSmithKline. All other authors declare no potential conflict of interest.

ETHICS STATEMENT

Scientific approval was obtained to use CPRD data by the Independent Scientific Advisory Committee (ISAC) (Protocol 17_194) and ethical approval from the London School of Hygiene & Tropical Medicine ethics committee.

ACKNOWLEDGEMENTS

J.T. is funded by a Medical Research Council PhD Studentship (MRC LID) grant MR/N013638/1. This work was supported by the Medical Research Council project grants MR/M013278/1 and MR/S01442X/1.

ORCID

John Tazare https://orcid.org/0000-0002-7194-2615

REFERENCES

1. Council of European Union. Council regulation (EU) no 1235/2010. 2010.
2. US FDA. *Guidance for Industry Postmarketing Studies and Clinical Trials — Implementation of Section 505(o)(3) of the Federal Food, Drug, and Cosmetic Act*. 2011. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/postmarketing-studies-and-clinical-trials-implementation-section-505o3-federal-food-drug-and-0.
3. Douglas IJ, Evans SJ, Hingorani AD, et al. Clopidogrel and interaction with proton pump inhibitors: comparison between cohort and within person study designs. *BMJ*. 2012;345:e4388. https://doi.org/10.1136/bmj.e4388.
4. Freemantle N, Marston L, Walters K, Wood J, Reynolds MR, Petersen I. Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ*. 2013;347:f6409. https://doi.org/10.1136/bmj.f6409.
5. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20(4):512-522. https://doi.org/10.1097/EDE.0b013e3181a663cc.
6. Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *J Clin Epidemiol*. 2018;10:771-788.
7. Suissa S, Dell'Aniello S, Ernst P. Concurrent use of long-acting bronchodilators in COPD and the risk of adverse cardiovascular events. *Eur Respir J*. 2017;49(5):1602245. https://doi.org/10.1183/13993003.02245-2016.
8. Suissa S, Dell'Aniello S, Ernst P. Long-acting bronchodilator initiation in COPD and the risk of adverse cardiopulmonary events: a population-based comparative safety study. *Chest*. 2017;151(1):60-67. https://doi.org/10.1016/j.chest.2016.08.001.
9. Toh S, Garcia Rodriguez LA, Hernan MA. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. *Pharmacoepidemiol Drug Saf*. 2011;20(8):849-857. https://doi.org/10.1002/pds.2152.
10. Jackson JW, Schmid I, Stuart EA. Propensity scores in pharmacoepidemiology: beyond the horizon. *Curr Epidemiol Reports*. 2017;4(4):271-280. https://doi.org/10.1007/s40471-017-0131-y.
11. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res*. 2011;46(3):399-424. https://doi.org/10.1080/00273171.2011.568786.
12. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from naive enthusiasm to intuitive understanding. *Stat Methods Med Res*. 2012;21(3):273-293. https://doi.org/10.1177/0962280210394483.
13. Williamson EJ, Forbes A. Introduction to propensity scores. *Respirology*. 2014;19(5):625-635. https://doi.org/10.1111/resp.12312.
14. World Health Organisation. International classification of diseases. Accessed December 10, 2019. www.who.int/classifications/icd/en/.
15. Bross IDJ. Spurious effects from an extraneous variable. *J Chronic Dis*. 1966;19:637-647.
16. Wyss R, Fireman B, Rassen JA, Schneeweiss S. Erratum: high-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2018;29(6):e63-e64.
17. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58(4):323-337. https://doi.org/10.1016/j.jclinepi.2004.10.012.
18. Franklin JM, Schneeweiss S. When and how can real world data analyses substitute for randomized controlled trials? *Clin Pharmacol Ther*. 2017;102(6):924-933. https://doi.org/10.1002/cpt.857.
19. Azoulay L, Eberg M, Benayoun S, Pollak M. 5alpha-Reductase inhibitors and the risk of cancer-related mortality in men with prostate cancer. *JAMA Oncol*. 2015;1(3):314-320. https://doi.org/10.1001/jamaoncol.2015.0387.
20. Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: clinical practice research Datalink (CPRD). *Int J Epidemiol*. 2015;44(3):827-836. https://doi.org/10.1093/ije/dyv098.
21. NHS Digital. Coding cross maps. https://isd.digital.nhs.uk/.
22. NHS Digital. BNF classification. https://bit.ly/2zGA1pR.
23. Wyss R, Schneeweiss S, van der Laan M, Lendle SD, Ju C, Franklin JM. Using super learner prediction modeling to improve high-dimensional propensity score estimation. *Epidemiology*. 2018;29(1):96-106.
24. Garbe E, Kloss S, Suling M, Pigeot I, Schneeweiss S. High-dimensional versus conventional propensity scores in a comparative effectiveness study of coxibs and reduced upper gastrointestinal complications. *Eur J Clin Pharmacol*. 2013;69(3):549-557. https://doi.org/10.1007/s00228-012-1334-2.
25. Whitaker HJ, Farrington CP, Spiessens B, Musonda P. Tutorial in biostatistics: the self-controlled case series method. *Stat Med*. 2005;0:1-31.
26. Holmes MV, Perel P, Shah T, Hingorani AD, Casas JP. CYP2C19 genotype, clopidogrel metabolism, platelet function, and cardiovascular events: a systematic review and meta-analysis. *JAMA*. 2011;306(24):2704-2714. https://doi.org/10.1001/jama.2011.1880.
27. Bhatt DL, Cryer BL, Contant CF, et al. Clopidogrel with or without omeprazole in coronary artery disease. *N Engl J Med*. 2010;363(20):1909-1917. https://doi.org/10.1056/NEJMoa1007964.
28. StataCorp. *Stata Statistical Software: Release*. Vol 14. College Station, TX: StataCorp LP; 2015.
29. Hallas J, Pottegard A. Performance of the high-dimensional propensity score in a Nordic healthcare model. *Basic Clin Pharmacol Toxicol*. 2017;120:312-317.
30. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol*. 2008;167(5):523-529. https://doi.org/10.1093/aje/kwm355.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Tazare J, Smeeth L, Evans SJW, Williamson E, Douglas IJ. Implementing high-dimensional propensity score principles to improve confounder adjustment in UK electronic health records. *Pharmacoepidemiol Drug Saf*. 2020;1–9. https://doi.org/10.1002/pds.5121